日本の学術出版物におけるオープン・サイテーションの分析

AUTHOR(S):

西岡, 千文; 亀田, 尭宙; 佐藤, 翔

CITATION:

西岡, 千文 ...[et al]. 日本の学術出版物におけるオープン・サイテーションの分析. 情報知識学会誌 2020, 30(1): 3-20

ISSUE DATE:

2020-02-29

URL:

http://hdl.handle.net/2433/254082

情報知識学会誌 2020 Vol. 30, No. 1

研究論文

日本の学術出版物における オープン・サイテーションの分析
Availability of Open Citation Data in Japanese Scholarly Literature

西岡 千文[1*] 亀田 尭宙[2] 佐藤 翔[3]
Chifumi NISHIOKA[1], Akihiro KAMEDA[2], Sho SATO[3]

[1] 京都大学附属図書館
Kyoto University Library
〒606-8501 京都市左京区吉田本町
E-mail: nishioka.chifumi.2c@kyoto-u.ac.jp
[2] 京都大学東南アジア地域研究研究所
Center for Southeast Asian Studies, Kyoto University
〒606-8501 京都市左京区吉田下阿達町 46
E-mail: kameda@cseas.kyoto-u.ac.jp
[3] 同志社大学免許資格課程センター
Center for License and Qualification, Doshisha University
〒302-8580 京都府京都市上京区新町通今出川上ル 同志社大学渓水館 315
E-mail: min2fly@slis.doshisha.ac.jp
* 連絡先著者 Corresponding Author

学術出版物のオープンアクセスが進展し，自由にアクセス可能な学術情報が蓄積されている．一方，引用データに関しては，機械可読なアクセスのオープン化が遅れている．このような状況を解決するために，I4OC（Initiative for Open Citations）が設立され，オープン・サイテーション，すなわち引用データのオープン化を推進している．本稿では，I4OC の取り組みによって公開された引用データを分析することで，日本におけるオープン・サイテーションの現状の把握を試みる．結果，世界の学術出版物において引用データをオープンにしている文献の割合は 24.22%であるのに対して，日本の学術出版物においては 18.86%であることが判明した．オープン・サイテーションを推進するための今後の課題として，過去の文献と人文学系分野の文献の引用データの組織化の支援が挙げられる．

In order to promote open citation of scholarly publications, I4OC (Initiative for Open Citations) has been launched by academic institutions and publishers. Open citation refers to citation data, which are structured, separable, and open. In this article, we analyze the citation data, which have been made available by I4OC, in order to understand the current status of open citation in Japan. The analysis reveals that while the percentage of publications with open citation data in the world is 24.22%, it is 18.86% in Japanese scholarly publications. In order to further promote open citation, it is necessary to support organizing and structuring citation data of past publications and publications in humanities.

キーワード: 引用データ，学術情報，オープンデータ，学術情報流通
citation data, scholarly information, open data, scholarly communication

1 はじめに

学術出版物の引用データは，研究評価，研究プロセスの理解，図書館の蔵書形成 [1]，学術論文推薦システムの構築 [2] 等，様々な目的で利用されている．しかし，学術出版物の引用データへの機械可読なアクセスの実現は遅れている．このような状況を解決するために，2017 年 4 月に国際的なイニシアティブである I4OC（Initiative for Open Citations）[3] が学術機関と出版社によって設立された．I4OC は，オープン・サイテーション，すなわち機械可読なフォーマットでの引用データのオープン化を推進することで，引用データの利用可能性を高めることを目的としている．具体的には，Crossref にメタデータとして登録された各文献の引用文献リスト，すなわち引用データ(注 1)をオープンにすることを推進している．Crossref DOI が付与された約 1 億件の出版物のうち，2018 年 9 月時点では 24.20%の文献の引用データがオープンになっている．

なお本稿では，Peroni らによるオープン・サイテーションの定義 [4] に倣って，以下の条件を満たす引用データを「オープン」であると判断する．この条件は I4OC でも使用されている．

- **構造的（Structured）**RDF 等の機械可読なフォーマットで表現されていること
- **分離可能（Separable）**引用データが記載されている引用元文献にアクセスしなくても，引用データを入手可能であること
- **オープン（Open）**無償でアクセス可能であり，再利用に際して制限がないこと

さらに，引用データにて引用元・引用先となっている文献は，下記の条件を満たさなければならない．

- **識別可能（Identifiable）**DOI 等の永久識別子によって識別可能であること
- **入手可能（Available）**識別子を利用することで文献の基本的なメタデータを入手可能であること

よって，オープン・サイテーションにおける「オープン」は，5 つ星オープンデータモデル [5] における 4 つ星あるいは 5 つ星であることを意味する．5 つ星オープンデータモデルとは，Tim Berners-Lee によって提唱されている，オープンデータを 5 段階で評価するスキームである．4 つ星は RDF など W3C が定める標準に沿ったフォーマットで公開されていることを，5 つ星は他のデータとリンクしていることを指す．I4OC は 5 つ星オープンデータを指向しており，I4OC の関連組織である OpenCitations[6] が RDF で表現された引用データのコーパスや SPARQL エンドポイントを提供している．

オープン・サイテーションにおける「オープン」がデータへの機械可読なアクセスを求めていることに対して，オープンアクセスにおける「オープン」は無償でアクセス可能なことのみが多くの場合で求められている．よって，オープンアクセスにおける「オープン」とオープン・サイテーションにおける「オープン」は異なる意味をもつ．文献がオープンアクセスで公開されていれば引用データにも無償でアクセス可能であるが，上述の条件を満たしていない引用データについては，オープンであると本稿では判断しない．

執筆時点では I4OC によるオープン・サイテーションのための取り組み開始から 2 年しか経過していないため，オープン・サイテーションが文献の発見可能性の向上 [7, 8] や引用回数の増加 [9, 7, 8] に貢献したといったオープンアクセスに与えられているような肯定的な評価を断定することはできない．しかし，これまでに I4OC によって公開された引用データは学術情報探索プラットフォーム Dimensions[10]，Wikidata，OpenAIRE[11]，LOC-DB[12] 等のプロジェクトや計量書誌学など，様々な場面での利活用が進んでいる [13]．また，電子図書館でのユーザの情報検索行動に関する研究 [14, 15] では，文献のページにおける引用文献・被引用文献の閲覧という情報探索行動は，クエリ検索など他の情報探索行動と比較して，探索によって得られた文献に肯定的な評価を示す行動（e.g., PDF ファイルの閲覧，書誌情報のエクスポート）をより多く引き起こすことが報告されている．よって，オープン・サイテーションが学術情報流通や研究評価に及ぼす影響は大きくなることが予想される．さらに，2018 年 9 月にオープンアクセス出版推進のために設立されたイニシアティブ Plan S では，学術雑誌への推奨事項として，「I4OC が規定する標準に沿った引用データへのアクセスの提供」を挙げている [16]．このことからも，学術情報流通においてオープン・サイテーションはさらに

(注 1) 本稿では，「文献の引用データ」は「文献の引用文献リスト」を意味する．

重要となってくるだろう．

本稿では，日本の学術出版物における引用データのオープン化の現状を分析する．「日本の学術出版物」とは，日本の著者により執筆された出版物ではなく，日本の出版者から公開された出版物を意味する．これは，学術出版物のオープンアクセスとは異なり，引用データをオープンにする主体は著者ではなく出版者であることが多いためである．分析にあたっては，OpenCitations から公開された引用データのコーパスを，JaLC メタデータ，Crossref メタデータ，unpaywall のデータセットとともに利用する．これらのデータセットを利用することで各文献の引用データのオープン化の状況を「オープン」，「クローズド」または「未整理・非存在」に分類する．引用データをオープン，クローズド，未整理・非存在にしている文献それぞれの，分類，出版プラットフォーム，出版年といった属性を把握することで，今後オープン・サイテーションを推進するにあたっての課題を把握することを目的とする．

本稿の構成を以下に示す．2 章では，関連研究を挙げることで，本稿と関連研究の差異を明らかにする．3 章では，引用データのオープン化の分析方法について詳述する．4 章では，出版物の分類，収録雑誌，出版年など様々な観点から，引用データのオープン化の状況について分析結果を示す．4 章から得られた結果に対する考察を 5 章に示す．6 章を本稿のむすびとする．

2 関連研究

本章は，引用データのオープン化とその分析についての関連研究を述べる．

近年，オープンサイエンスの潮流を背景として，学術情報分野においてもオープンデータが進展している [17]．学術情報分野のオープンデータとして，Microsoft Academic Graph[18]，Springer Nature SciGraph[19] や，特定の分野に限定した DBLP（情報学）[20]，APS Data Sets（物理学）[21]，AMiner（情報学）[22, 23] 等が公開されている．これらのデータセットには，文献の書誌情報や著者情報，引用データといった様々なデータが収録されている．

特に Microsoft Academic Graph は，あらゆる分野の学術情報を収録対象としていることから，多くの分析がなされている．Herrmannova ら [24] は，2016 年 2 月の時点で，Microsoft Academic Graph において，収録されている文献のうち 23.7%の文献のみに引用データが登録されていることを報告している．Harzing ら [25] は，Microsoft Academic Graph，Google Scholar，Scopus，Web of Science の各データセット・データベースから計算される被引用回数の比較を実施している．Haunschild ら [26] は，Microsoft Academic Graph と Web of Science における引用データを比較している．両者において，引用文献数の相関はほぼ観察されなかった．Microsoft Academic Graph API では各文献の取得可能な引用文献件数が最大 50 件であるため，引用文献件数が 50 件未満であるものに限定した場合，相関が観察された．

本稿では，オープン・サイテーションの現状を Microsoft Academic Graph ではなく，COCI（OpenCitations Index of Crossref Open DOI-to-DOI references）[27] を利用して分析する．COCI は，I4OC の活動によってオープンとなった文献の全引用データを収録しているコーパスであり，I4OC の関連組織である OpenCitations によって作成・公開されている．Microsoft Academic Graph の各文献の引用データの生成過程が不明瞭であることに対して，COCI は出版者によって登録された引用データが収録されていることから信頼性が高い．COCI については，3.1.1 項で詳述する．

COCI を利用したオープン・サイテーションの現状分析として，Heibi らによるもの [28] が挙げられる．COCI において各引用は引用文献の DOI と被引用文献の DOI のペアとして表現され，COCI は Crossref に登録されている全引用のリストである．Heibi らは，各引用に注目して，各引用のオープン化の状況を，引用文献の種別・出版者別に分析している．本稿では，各引用ではなく各文献に注目して，オープン・サイテーションの状況を分析した．Crossref において引用データの公開・非公開の設定は，引用単位で行うことはできず，文献単位で設定される．よって本稿では，文献に注目した分析を実施することで，引用データをオープン，クローズド，未整理・非存在にしている文献それぞれの分類や出版年といった特徴を把握する．

筆者らの先行研究 [29] では，COCI を利用して日本におけるオープン・サイテーションの現状を分析した．先行研究では，各文献の引用データのオープン化の状況を「オープン」あるいは「その他」とラベル付けして分類した．本稿では先行研究で「その

他」と分類されていた文献をさらに,「クローズド」と「未整理・非存在」に分類（3.2 節参照）し，より詳細なオープン・サイテーションの現状の把握を可能とした．また，先行研究ではなされなかった，出版プラットフォーム別のオープン・サイテーションの分析を実施している．さらに分析に利用したデータセットが異なる．先行研究時には Crossref メタデータのデータセットを取得できなかったことから，unpaywall データセットを利用しており，文献のメタデータならびにオープンアクセスの状況を unpaywall より取得している．本稿では unpaywall のデータセットとともに，Crossref メタデータのデータセットを利用しており，Crossref メタデータのフィールド `reference-count` により，先述のより詳細な文献のラベル付けが可能となっている（3.2 節参照）．Crossref メタデータのデータセットならびに unpaywall データセットについては，3.1.2 項で詳述する．

3 手法

本章は，日本の学術出版物における引用データのオープン化の分析方法について述べる．3.1 節では，分析に利用したデータセットについて紹介する．3.2 節では，各文献の引用データのオープン化状況の判断方法について詳述する．3.3 節では，各文献の分類の特定する手法について述べる．

3.1 データセット

本節は分析に利用したデータセットについて記す．3.1.1 項では引用データについて，3.1.2 項では本稿での分析対象となる日本の学術出版物についてそれぞれ述べる．

3.1.1 引用データ

引用データとして，I4OC の活動によって公開されたデータセットを利用する．I4OC は Crossref で各文献の引用データ，すなわち各文献の引用文献リストを公開することを出版者に推奨している．Crossref では，各出版者は各文献の引用データの公開状況を「公開（open）」「限定（limited）」「非公開（closed）」のうちいずれかに設定している [30].「公開（open）」と設定された場合は，Crossref API から文献の引用文献リストを取得可能である.「限定（limited）」は，Crossref Metadata APIs Plus Service[31] の購読者とメタデータの登録を行った出版者のみが引用文献リストを取得可能である.「非公開（closed）」は，メタデータの登録を行った出版者以外には非公開であることを示す．I4OC の活動によって公開された引用データは，I4OC の関連組織である OpenCitations によって，RDF，csv 等の機械可読なフォーマットでデータセットとして公開されている．OpenCitations は Crossref Metadata APIs Plus Service の購読者であり,「限定（limited）」で公開されている引用データの再配布に制限はないため，引用文献リストの公開設定が「公開（open）」あるいは「限定（limited）」である文献の引用データを配布している [28]．本稿では，COCI（OpenCitations Index of Crossref Open DOI-to-DOI references）（2018 年 11 月 22 日版）[32, 27] を引用データとして分析に利用する．データセットの名称のとおり，引用元となる文献については Crossref DOI をもつ文献のみが対象となっている．引用先となる文献については Crossref DOI を含むあらゆる DOI をもつ文献が対象である．表 1 にデータセット概要を示す．表 1 にあるとおり，データセットには 46,530,436 件の文献間の引用(注 2) が 449,842,375 件記述されている．なお，これらの数値はウェブサイトに記述されている数値 [33] と異なる．本稿では著者らが精査した結果である表 1 の数値を利用する．

なお，オープンになっている引用データとして Microsoft Academic Graph[18] も挙げられるが，引用データの生成過程が不明瞭といった理由から，本稿では COCI データセットを利用して分析を実施する（2 章参照）.

表 1: COCI データセット（2018 年 11 月 22 日版）概要.

	件数
引用文献として収録されている文献	24,182,977
被引用文献として収録されている文献	38,481,195
収録されている文献	46,530,436
収録引用数	449,842,375

(注 2) 引用文献の DOI と被引用文献の DOI のペアによって表現される.

3.1.2 日本の学術出版物

1 章で述べたとおり，本稿における「日本の学術出版物」とは日本の出版者から公開された出版物を意図する．本稿では日本の学術出版物を「ジャパンリンクセンター（JaLC）によって付与された DOI をもつ出版物」であると仮定し，2018 年 9 月 7 日版の JaLC メタデータ [34] を利用する．JaLC メタデータには，JaLC によって付与された DOI をもつ文献が 6,370,356 件収録されている．しかし前節で述べたとおり，COCI で引用文献として収録されている文献（引用文献リストを公開している文献）は，Crossref DOI をもつ文献のみである(注 3)．よって，JaLC DOI が付与された学術出版物の引用データは COCI に収録されていない．

そのため本稿では，JaLC 経由で登録された Crossref DOI を保有する文献を分析対象とする．JaLC は，JaLC DOI，Crossref DOI ならびに DataCite DOI を文献に付与することが可能であることから，JaLC メタデータは Crossref DOI をもつ文献も収録されている．JaLC 経由で登録された Crossref DOI をもつ文献は，JaLC メタデータに収録されている文献の集合と Crossref メタデータに収録されている文献の集合の積集合から特定する．Crossref メタデータとして，2018 年 9 月に収集・作成されたデータセット [35] を利用する．データセットは，Crossref DOI が付与されている 99,874,789 件の文献の書誌情報等のメタデータを収録しており，2018 年 9 月 5 日の Crossref のデータベースの状態を反映している．データセットの収集に使用されたスクリプトは公開されている [36]．

さらに本稿では，各文献のオープンアクセス状況に応じたオープン・サイテーションの状況についても調査する．そのため，2018 年 9 月 24 日版の unpaywall データセット [37] を利用する．unpaywall はあらゆる文献のウェブ上で合法的にオープンアクセスとなっている版を探索・提供するウェブブラウザの拡張機能であり，Impactstory から提供されている．拡張機能で利用されているデータベースのスナップショットはデータセットとして公開されており，Crossref DOI を保有する全文献の DOI ならびにオープンアクセスに関する情報等を収録している．本稿で使用したデータセットには，99,940,229 件の文献が収録されていた．

本稿では，JaLC メタデータ，Crossref メタデータ，unpaywall データセット全てに収録されている文献を調査対象とする．Crossref メタデータと unpaywall データセットの両方に含まれる文献は，99,848,571 件である．そのうち，JaLC メタデータに収録されている文献は 2,049,891 件であり，JaLC メタデータ全文献のうち 32.18%を占める．

3.2 各文献のオープン・サイテーション状況のラベル付け

各文献のオープン・サイテーション状況，すなわち引用文献リストの公開状況については，下記のうちいずれかのラベルを割り当てる．

- **オープン**: 「COCI データセットで引用文献として文献が収録されている」場合，引用文献リストをオープンにしている文献とする．
- **クローズド**: 「COCI データセットで引用文献として文献が収録されていない」かつ「当該文献の Crossref メタデータのフィールド `reference-count` が 1 以上である」場合，引用文献を非公開にしている文献と判定する．`reference-count` は，文献の引用文献件数を記録するフィールドである．引用文献が 1 件以上あるにも関わらず COCI データセットに引用文献として文献が収録されていないということは，出版者が引用文献リストを非公開に設定していることを意味する．
- **未整理・非存在**: 「上記のいずれにも該当しない」場合，当ラベルを割り当てる．このラベルに該当する状態を下記に挙げる．
  - **未整理** 文献の引用文献リストが組織化されておらず，オープン・サイテーションの条件を満たす引用データとして整理されていない．引用文献が存在するにも関わらず，予算等のリソース不足といった理由から，引用文献が Crossref に登録されていない状態を指す．
  - **非存在** 文献に引用文献が存在しない．例として，雑誌の編集後記や目次が挙げられる．

(注 3) 一件の例外が観察された．文献（`https://doi.org/10.1589/jpts.30.`）は，JaLC DOI が付与されているが，COCI にて引用元の文献として収録されていた．

Crossref メタデータでは，上記のいずれの状態でもフィールド `reference-count` に 0 が与えられるため，状態を区別することが不可能である．「引用文献が存在しない」ことの明示を可能にするために，RDF における非存在の記述 [38] 等を勘案した検討が必要となる．

3.3 文献の分類の特定

本稿では，文献の分類ごとに引用データのオープン化の状況を分析する（4.3 節）．このことから，本節では文献の分類を特定する手法について述べる．文献の分類は，国立国会図書館が提供する雑誌記事索引採録誌一覧 [39] を利用することで特定した．まず各文献を ISSN によって，雑誌記事索引採録誌一覧に収録されている雑誌に紐付ける．各文献の ISSN は Crossref メタデータより取得した．雑誌記事索引採録誌一覧には 23,910 誌が収録されており，そのうち 17,897 誌の ISSN が記載されている．分類として，雑誌記事索引採録誌一覧で各雑誌に付与されている国立国会図書館分類表の分類 [40] を利用する．各雑誌には逐次刊行物 [41] に属する分類が付与されている．特に，本稿では雑誌の内容に関する ZA～ZS の分類に着目する．分類が取得できなかった場合は，「不明」という分類に割り当てた．なお，分類が複数存在する文献については，1 を与えられた分類数で割り，各分類に割り当てる．

4 結果

本章では，オープン・サイテーションの現状についての分析結果について述べる．4.1 節では，世界の学術出版物におけるオープン・サイテーションの現状との比較を示す．続いて，オープンアクセス状況（4.2 節），分類（4.3 節），収録雑誌（4.4 節），出版プラットフォーム（4.5 節），出版年（4.6 節）といった観点から，日本のオープン・サイテーションの状況について報告する．4.7 節では，紀要に焦点を当てて，オープン・サイテーションの状況を分析する．

4.1 世界の学術出版物におけるオープン・サイテーションの現状との比較

本節では，日本の学術出版物におけるオープン・サイテーションの現状を，世界の学術出版物における現状と比較する．表 2 に，日本の学術出版物・世界の学術出版物それぞれについて，オープン・サイテーションの現状を示す．世界の学術出版物と比較すると日本の学術出版物では，引用データをオープン・クローズドにしている文献の割合が低く，未整理・非存在である文献の割合が高い．分析対象となっている日本の学術出版物のうち 99.99%は Crossref メタデータにおける文献種別が `journal-article` であることから，これらの文献のうち多くが何かしらの文献を引用していると考えらる．よって，これらの文献は，引用文献が存在しない文献というよりも，引用データが未整理である文献であるといえよう．これらの文献の引用データを整理することが，今後の課題である．

4.2 オープンアクセス状況

本節では，オープンアクセスである文献とそうでない文献のあいだに，オープン・サイテーションの状況に差異が存在するか調査する．3.1.2 節で述べたとおり，unpaywall データセットから各文献のオープンアクセス状況を取得する．

オープンアクセス状況別のオープン・サイテーションの状況の分析結果を表 3 に示す．比較のため，世界の出版物における分析結果も示している．表 3 におけるオープンアクセスは，ゴールドオープンアクセス，グリーンオープンアクセスなどあらゆるオープンアクセスを含む．オープンアクセス状況についてであるが，日本の学術出版物では，2,049,891 件のうち 1,696,636 件（82.77%）がいずれかの方法でオープンアクセスとなっていることが判明した．対して，世界の学術出版物では，99,848,571 件のうち 24,961,752 件（25.00%）がオープンアクセスになっている．25.00%という数値は，Piwowar らのオープンアクセスの現状の調査 [9] で報告された数値と概ね一致する．海外と比較すると，日本の学術出版物は，商業出版社ではなく学協会から出版されるものが多いため，オープンアクセスの割合が高くなっているといえる．日本の学術出版物におけるオープン・サイテーションの状況についてであるが，若干

表 2: 日本の学術出版物と世界の学術出版物におけるオープン・サイテーションの現状.

	全文献件数	オープン		クローズド		未整理・非存在	
日本 [1]	2,049,891	386,632	(18.86)	148,130	(7.23)	1,515,129	(73.91)
世界 [2]	99,848,571	24,178,446	(24.22)	16,589,545	(16.61)	59,080,580	(59.17)

それぞれにおいて，3.2 節で述べた各ラベルに分類される文献件数と全文献件数におけるその割合（括弧内）を百分率で示す.

[1] JaLC メタデータ，Crossref メタデータ，unpaywall データセット全てに収録されている文献（≈JaLC 経由で登録された Crossref DOI をもつ文献）.
[2] Crossref メタデータかつ unpaywall データセットに収録されている文献.

オープンアクセスでない文献で引用データのオープン化が進んでいるものの，大差はない．クローズドや未整理・非存在に属する文献の割合についても，オープンアクセスである文献とオープンアクセスでない文献のあいだにほとんど差はない．世界の学術出版物においては，オープンアクセスである文献において，引用データのオープン化がより進んでいるが，差は大きいとはいえない．クローズドについては，オープンアクセスでない文献で割合がより高くなっている．

表 3 の調査におけるオープンアクセスは，あらゆるオープンアクセスを対象としていた．グリーンオープンアクセスにおいては，出版者版へのアクセスには購読料が必要で，引用データもオープンにされていないということが考えられる．上記の可能性を除くため，オープンアクセスジャーナル・非オープンアクセスジャーナルにおけるオープン・サイテーションの状況を調査した．unpaywall データセットには，`journal_is_oa` という「文献がオープンアクセスジャーナルで出版されているか」，`journal_is_in_doaj` という「文献が Directory of Open Access Journals（DOAJ）[42] の収録雑誌で出版されているか」を記録するフィールドがそれぞれ存在する [43]．`journal_is_oa` は開発中 [43] であることから，分析では `journal_is_in_doaj` を利用する．DOAJ はオープンアクセスジャーナルを，「全コンテンツに無償でアクセス可能である」かつ「査読・編集によって高品質基準を満たしている」学術雑誌と定義しており，これらの要件を満たす学術雑誌のリストを公開している．学術雑誌が DOAJ に掲載されるためには，DOAJ が定める要件を満たさなければならない．このような学術雑誌では，引用データが整理されやすいと考える．結果を表 4 に示す．日本の学術出版物において DOAJ に収録されているものは，3,760 件（0.18%）と非常に少ない．しかし，そのうちの約半数である 49.81%が引用データをオープンにしている．一方，26.38%に相当する 992 件が引用データを非公開にしている．

世界の学術出版物においても，DOAJ に収録されているオープンアクセスジャーナルの文献のうち，約半数である 48.99%が引用データをオープンにしている．非公開にしている文献は 8.47%で日本の学術出版物における割合よりも低く，文献がオープン・サイテーションではない理由として引用データが組織化されていないことが挙げられる．海外では大学が発行するオープンアクセスジャーナルが DOAJ に登録されているケースも多く，これらの出版者で引用データを整理するためのリソースがないことが，引用データが組織化されていない理由として挙げられる．また，DOAJ は英語以外のオープンアクセスジャーナルも多く収録している．引用データが未整理・非存在である文献が収録されているオープンアクセスジャーナルのタイトルの言語を抽出したところ，6,813 誌のうち英語のタイトルをもつ雑誌は 3,595 誌（52.77%）に留まる．タイトルからの言語抽出には，langdetect[44] を利用した．英語を除くと，ポルトガル語，スペイン語，インドネシア語の順に多い．英語と比較すると，これらの言語の文献では引用文献リストの抽出技術 [45, 46] の開発があまりなされていないことも，引用データが整理されていない理由として挙げられる．

4.3 分類

本節では，オープン・サイテーションの現状を文献の分類別に探る．各文献の分類の特定手法は，3.3 節で述べた．

表 5 に分析結果を示す．化学・化学工業の文献で，引用データのオープン化が特に進展していることがわかる．26.12%という数値は，表 2 に示されている世界の学術出版物における引用データをオープンにしている文献の割合 24.20%より高い．その他，生

表 3: オープンアクセス（OA）状況別のオープン・サイテーションの状況.

		全文献件数	オープン		クローズド		未整理・非存在	
日本	OA 文献	1,696,636	316,810	(18.67)	124,015	(7.31)	1,255,811	(74.02)
	非 OA 文献	353,255	69,822	(19.77)	24,115	(6.83)	259,318	(73.41)
世界	OA 文献	24,961,752	7,077,997	(28.36)	2,391,627	(9.58)	15,492,128	(62.06)
	非 OA 文献	74,886,819	17,100,449	(22.84)	14,197,918	(18.96)	43,588,452	(58.21)

それぞれにおいて，3.2 節で述べた各ラベルに分類される文献件数と全文献件数におけるその割合（括弧内）を百分率で示す.

表 4: DOAJ 収録雑誌・DOAJ 非収録雑誌におけるオープン・サイテーションの状況.

		全文献件数	オープン		クローズド		未整理・非存在	
日本	DOAJ 収録雑誌文献	3,760	1,837	(49.81)	992	(26.38)	895	(23.80)
	DOAJ 非収録雑誌文献	2,046,131	384,759	(18.80)	147,138	(7.19)	1,514,234	(74.00)
世界	DOAJ 収録雑誌文献	3,227,017	1,580,857	(48.99)	273,392	(8.47)	1,372,768	(42.54)
	DOAJ 非収録雑誌文献	96,621,554	22,597,589	(23.39)	16,316,153	(16.89)	57,707,812	(59.72)

それぞれにおいて，3.2 節で述べた各ラベルに分類される文献件数と全文献件数におけるその割合（括弧内）を百分率で示す.

表 5: 分類別のオープン・サイテーションの状況.

	全文献件数	オープン		クローズド		未整理・非存在	
政治・法律・行政	4,735.00	99.50	(2.10)	55.00	(1.16)	4,580.50	(96.74)
経済	10,179.00	1,534.00	(15.07)	1,675.00	(16.46)	6,970.00	(68.47)
社会・労働	10,758.50	1,446.00	(13.44)	1,367.00	(12.71)	7,945.50	(73.85)
教育	17,621.00	3,234.00	(18.35)	1,991.50	(11.30)	12,395.50	(70.35)
歴史・地理	14,740.00	192.00	(1.30)	990.00	(6.72)	13,558.00	(91.98)
哲学・宗教	15,104.00	1.00	(0.01)	124.00	(0.82)	14,979.00	(99.17)
芸術・言語・文学	6,162.00	595.00	(9.66)	605.00	(9.82)	4,962.00	(80.53)
科学技術	164,957.75	28,172.00	(17.08)	11,250.25	(6.82)	125,535.50	(76.10)
建設工学・機械工学 [1]	248,957.25	34,152.50	(13.72)	26,390.75	(10.60)	188,414.00	(75.68)
化学・化学工業 [2]	364,817.75	95,301.50	(26.12)	27,415.25	(7.51	242,101.00	(66.36)
生物学・農林水産	192,386.00	35,434.00	(18.42)	17,268.00	(8.98)	139,684.00	(72.61)
心理学・医学・薬学 [3]	407,761.75	68,203.50	(16.73)	21,191.25	(5.20)	318,367.00	(78.08)
分類不明	591,711.00	118,267.00	(19.99)	37,807.00	(6.39)	435,637.00	(73.62)
合計	2,049,891.00	386,632.00	(18.86)	148,130.00	(7.23)	1,515,129.00	(73.91)

それぞれにおいて，3.2 節で述べた各ラベルに分類される文献件数と全文献件数におけるその割合（括弧内）を百分率で示す.

[1] 建設工学・建設業，機械工学・工業，運輸工学，電気工学・電気機械工業，原子力工学・工業が含まれる.
[2] 化学・化学工業，繊維工学・工業，食品工学・工業，金属工学・鉱山工学，印写工学，その他の工学・工業が含まれる.
[3] 人類学，心理学，医学，薬学が含まれる.

物学，科学技術一般など，STM（科学・医学・工学）分野において，引用データのオープン化が進展していることがわかる．対して，歴史・地理，哲学・宗教，芸術など人文学系分野においては，引用データのオープン化は遅れている．STM 分野と比較するとこれらの分野では，未整理・非存在に分類されている文献の割合が高いことから，引用データの組織化の支援が求められる．

4.4 収録雑誌

文献のメタデータの付与・登録方法は出版者や編集者によって定められており，各雑誌の収録文献のあいだでは概ね統一されている．そのため，引用文献の Crossref メタデータへの登録やそれらの公開については，出版者や編集者の意向が反映される．よって，引用データのオープン化の状況は，収録雑誌によって大きく異なると考えられる．本節では，収録

雑誌別にオープン・サイテーションの状況を分析する．日本の学術出版物 2,049,891 件の雑誌の異なり数は 1,267 誌である．文献の収録雑誌は，Crossref メタデータより取得している．各雑誌ごとに，オープン，クローズドな引用データの割合をそれぞれ算出する．

結果の分布を図 1 に表す．横軸は百分率での割合を，縦軸はその割合に属する雑誌件数を示す．例えば，収録されている 10〜20%の文献の引用データをオープンにしている雑誌は約 100 誌あることがわかる．図 1 より，ほとんどの雑誌で，収録されている文献のうち 10%以下の文献で引用データをオープン，クローズドにしていることがわかる．1,267 誌のうち，904 誌で 1 件以上の文献の引用データをオープンにしていることが判明した．

どのような雑誌で引用データのオープン化が進展しているか考察するために，収録文献が 100 件以上ある雑誌のうち引用データをオープンにしている割合が高い雑誌上位 10 件を表 6 に示す．全ての雑誌が STM 分野に属する雑誌であり，いずれも英文誌である．これらの雑誌の多くは Web of Science Core Collection に収録されていてインパクトファクターが付与されていることから，国際的に認知されているといえる．また，日本の学術出版物の多くが J-STAGE や機関リポジトリを出版プラットフォームとして利用していることに対して，これらの雑誌のうち 5 誌が出版プラットフォームとして Nature, Wiley Online Library といった商業出版社のプラットフォームや，海外でも広く利用されている Atypon 社の出版プラットフォームを利用している．このことから，出版プラットフォーム等での慣習が，引用データの登録・オープン化に影響を与えていることが考えられる．

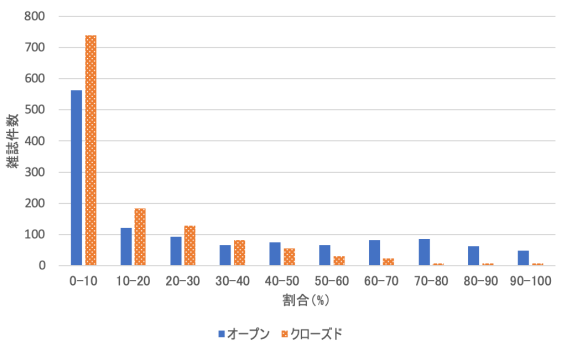

図 1: 雑誌別引用データをオープンにしている割合の分布．

表 6: 引用データをオープンにしている割合が高い雑誌上位 10 誌．

雑誌名	割合（%）
BIOPHYSICS	100.00
Polymer Journal	99.58
MICROBIOLOGY and IMMUNOLOGY	99.55
Chemistry Letters	99.08
MATERIALS TRANSACTIONS	98.41
Circulation Journal	98.33
Journal of the Meteorological Society of Japan	98.33
Microbes and Environments	97.94
Materials Transactions, JIM	97.88
Journal of the Physical Society of Japan	97.64

4.5 出版プラットフォーム

前節において，各文献の引用データの登録・オープン化が出版プラットフォームでの慣習に影響されている可能性を指摘した．このことを調査するために，本節では，出版プラットフォームごとにオープン・サイテーションの状況を探る．

出版プラットフォームは，DOI のリダイレクト先 URL の Top Level Domain（TLD）と Second-Level Domain から特定した．すなわち，リダイレクト先 URL の TLD と Second-Level Domain が同一の文献は，同一の出版プラットフォームを利用しているとする．本稿での TLD と Second-Level Domain は，Internet Corporation for Assigned Names and Numbers（ICANN）[47] による定義に従う．リダイレクト先 URL の TLD と Second-Level Domain は，Python のモジュール tldextract[48] を利用して抽出する．なお，複数の URL を経由して最終的なランディングページに到達する DOI も存在するが，出版者サイトへの過度なアクセスを避けるため，DOI の第一リダイレクト先 URL によって集計を行っている．そのため，得られた TLD と Second-Level Domain が，最終的なランディングページのものと異なる場合もある．

表 7 は，最も利用されている出版プラットフォーム上位 10 件と各出版プラットフォームにおける文献のオープン・サイテーションの状況を示す．最も使用されている出版プラットフォームは jst.go.jp であるが，確認したかぎりいずれも J-STAGE を示す．よって，JaLC 経由で付与された CrossRef DOI を

もつ文献のうち，94.37%が J-STAGE で出版されている．J-STAGE においては，表 2 で報告された全文献における数値と比較すると，引用データがオープン，クローズドになっている文献の割合が低く，未整理・非存在に分類される文献が多い．2 位に位置する csj.jp は，公益社団法人日本化学会の出版プラットフォームを示しており，表 6 にもある Chemistry Letters や Bulletin of the Chemical Society of Japan といった英語の論文誌が出版されている．この出版プラットフォームでの引用データをオープンにしている文献の割合の高さが，表 5 で示されている化学・化学工業における引用データをオープンにしている文献の割合の高さにつながっている．この出版プラットフォームは Atypon 社のデジタルコンテンツシステムによって構築されており，日本物理学会の出版プラットフォーム（jps.jp）も同様のプラットフォームを導入している．いずれの出版プラットフォームにおいても，引用データをオープンにしている文献の割合は 97%以上と非常に高く，デジタルコンテンツシステムによる影響を受けていることが推察される．非営利の数学分野の出版プラットフォームである projecteuclid.org では，引用データをオープンにしている文献，クローズドにしている文献の割合がともに 0%であり，引用データの組織化が進んでいない．一方，商業出版者の出版プラットフォームである tandfonline.com（Taylor & Francis），nature.com（nature），elsevier.com（Elsevier）では，引用データをオープンにしている文献の割合には差があるものの，いずれも未整理・非存在に分類される文献は他の出版プラットフォームと比較すると少ない．

4.6 出版年

引用データは著作権などで保護されているという考えもあるため，出版社が定めるエンバーゴ期間にはオープンにできないということも考えられる．本節では，文献の出版年別にオープン・サイテーションの状況を調査する．各文献の出版年は Crossref メタデータより取得した．

図 2 に，日本の学術出版物における，出版年ごとの全文献件数，引用データをオープン，クローズドにしている文献件数とその割合を示す．対象とする出版年は 1990〜2018 年であり，この期間に出版された文献は全文献件数の 57.05%を占める．なお，本稿で利用しているデータセットは 2018 年 9 月に公開されたものであるため，2018 年に出版された文献は他の年と比較すると少ない．直近の 10 年間に出版された文献に注目すると，概ね 40%以上の文献の引用データがオープンになっていることがわかる．しかし，引用データをクローズドにしている文献の件数・割合も増加傾向が観察され，近年では出版年によってばらつきが見られるものの，30%前後で推移している．過去の文献では，引用データをオープン，クローズドにしている文献の割合はともに低く，引用データが組織化されていないことが推察できる．一般的に新しい文献と比較すると過去の文献が閲覧される回数は少ない [9] が，研究の検証可能性や学術の発展の追跡可能性の向上のためにも，過去の文献の引用データの組織化・オープン化が必要となる．

日本の状況を世界の状況と比較するために，図 3 に世界の学術出版物における出版年ごとの全文献件数，引用データをオープン，クローズドにしている文献件数とその割合を示す．図 2 と比較すると，世界の学術出版物では出版年がオープン，クローズドの割合に与える影響は少ない．引用データをオープンにしている文献についてであるが，2000 年以前は，日本の学術出版物における割合より高い 18%前後で推移している．その後，2003 年以降は，2013 年と 2017 年を除く出版年で，日本の学術出版物における引用データをオープンにしている文献の割合は世界の学術出版物と比較して高い．よって，近年に注目すると，日本の学術出版物におけるオープン・サイテーションは海外と比較して進んでいるといえる．世界の学術出版物における近年の引用文献のオープン化の割合が低い理由の一つとして，大手商業出版社による影響が挙げられる．Heibi らの調査 [28] では，Elsevier が出版元である Crossref DOI が付与された 16,773,716 件の文献のうち，11,020,314 件が引用文献をメタデータに記述している．しかし，それら全件はクローズドである．それらがオープンになることで，各出版年におけるオープン化の割合は 10%以上増加する．

4.7 紀要

大学などの研究機関が発行する学術雑誌である紀要は，機関リポジトリがプライマリな公開場所として利用されていることが多い [49]．J-STAGE や商

表 7: 最も利用されている出版プラットフォーム上位 10 件のオープン・サイテーションの状況.

	全文献件数	オープン		クローズド		未整理・非存在	
jst.go.jp	1,916,859	294,593	(15.37)	135,024	(7.04)	1,487,242	(77.59)
csj.jp	63,347	61,963	(97.82)	1,046	(1.65)	338	(0.53)
japanlinkcenter.org[1]	17,540	10,696	(60.98)	3,753	(21.40)	3,091	(17.62)
tandfonline.com	13,000	6,172	(47.48)	4,964	(38.18)	1,864	(14.34)
projecteuclid.org	11,360	0	(0.00)	0	(0.00)	11,360	(100.00)
不明	5,752	681	(11.84)	951	(16.53)	4,120	(71.63)
nature.com	4,827	4,813	(99.71)	10	(0.21)	4	(0.08)
oup.com	2,714	795	(29.29)	19	(0.70)	1,900	(70.01)
elsevier.com	2,711	843	(31.10)	1,822	(67.21)	46	(1.70)
jps.jp	2,505	2,443	(97.52)	3	(0.12)	59	(2.36)
合計	2,049,891.00	386,632.00	(18.86)	148,130.00	(7.23)	1,515,129.00	(73.91)

それぞれにおいて，3.2 節で述べた各ラベルに分類される文献件数と全文献件数におけるその割合（括弧内）を百分率で示す.
[1] 多くの場合，最終的なランディングページの TLD と Second-Level Domain は jst.go.jp である.

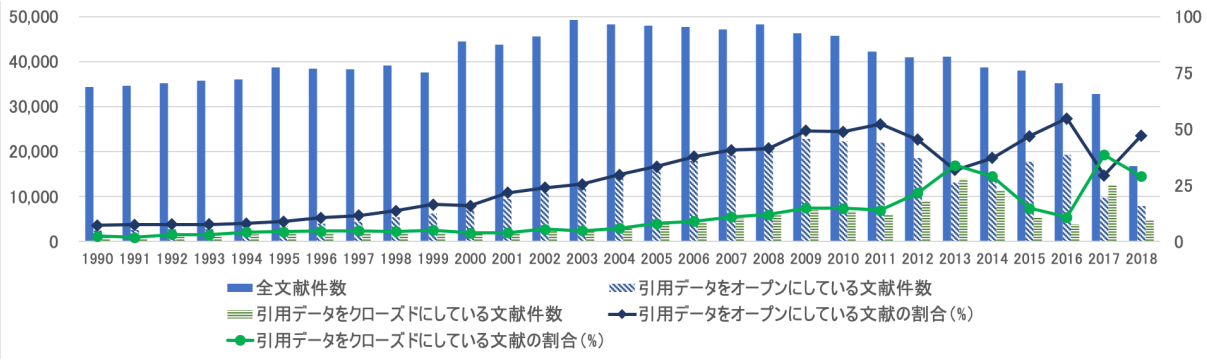

図 2: 日本の学術出版物における出版年ごとのオープン・サイテーションの状況.

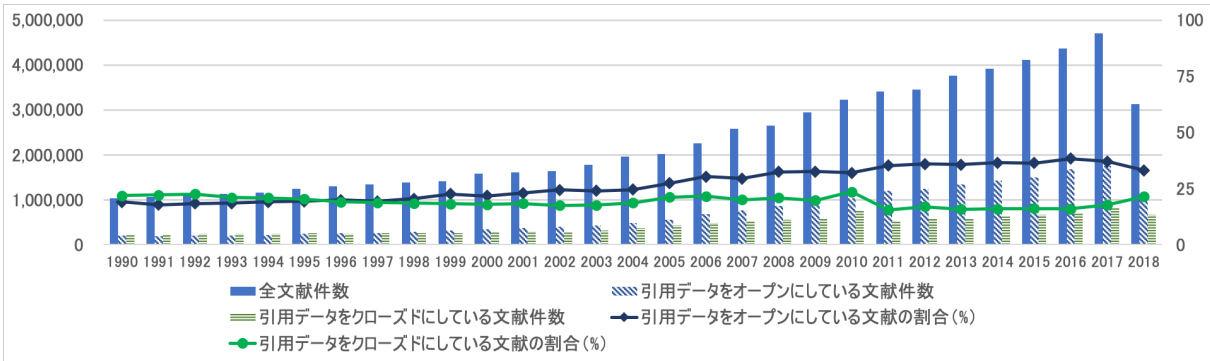

図 3: 世界の学術出版物における出版年ごとのオープン・サイテーションの状況.

業出版社のプラットフォームでは各文献のページにて引用文献や被引用文献へのリンクが付与されている場合が多く，文献に述べられている科学的知見の検証可能性の向上に寄与している．ユーザの情報検索行動に関する研究 [14, 15] では，これらのリンクが重要であることが示されている．対して，多くの機関リポジトリでは引用文献・被引用文献へのリンクが付与されていない場合が多く，同様の機能が望まれる．このような機能を検討するにあたっては文献の引用データがオープンになっていることが望ましい．このことから，紀要に収録されている文献に焦点を当てて，オープン・サイテーションの状況を分

析する．本節では，分析対象となっている文献から紀要を二通りの方法で特定して，分析を実施した．それぞれの特定方法における結果を報告する．

国立国会図書館分類で紀要に分類される文献　紀要に収録されている文献として，国立国会図書館分類 ZV1（紀要）をもつ文献を抽出する．雑誌記事索引採録誌一覧では，掲載されている 23,910 誌のうち 3,029 誌に ZV1 の分類が与えられている．4.3 節と同様に，ISSN で紐付けを行った結果，6,533 件の文献が ZV1 の分類が割り当てられていることがわかった．6,533 件の文献は，日本學士院紀要，Proceedings of the Japan Academy，Synthesiology，社会技術研究論文集，人間環境学研究のいずれかに収録されている．そのうち 2,131 件（32.62%）の文献が引用データをオープンにしており，499 件（7.64%）がクローズドにしている．3,903 件（59.74%）が未整理・非存在に分類される．前述の 5 誌いずれにおいても，1 件以上の引用データをそれぞれオープン，クローズドにしていることが判明した．32.62%という数値は，表 2 で報告されている日本の学術出版物，世界の学術出版物についての数値と比較しても高いといえるが，標本数が少ないため「紀要において引用データのオープン化が進展している」といった結論を導き出すことはできない．また，分析対象となった 5 誌はいずれも J-STAGE を出版プラットフォームとして利用しており，引用文献リストが存在する文献については，各引用文献へのリンクが与えられている．

大学が出版者となっている文献　文献の出版者名に「University」または「College」が入っている文献を抽出し，引用データのオープン化の状況を探った．26 件の出版者が該当したが，そのうち 18 件は出版社（e.g., オックスフォード大学出版局）・学会等であり，機関リポジトリを出版プラットフォームとして利用していない．これらを除いた 8 件は，機関リポジトリを利用して紀要を出版している．これらの 8 件の出版者が出版する雑誌は，前段落で挙げた 5 誌いずれも含まない．8 件の出版者は計 6,572 件の文献を出版しており，そのうち 2,215 件（33.70%）の文献が引用データをオープンにしている．1,228 件（18.69%）が引用データをクローズドにしており，3,129 件（47.61%）が未整理・非存在に分類される．紀要はいずれもオープンアクセスであるが，オープンアクセスであるにも関わらず引用データをクローズドにしている文献が一定数存在していることから，オープン・サイテーションの認知の向上が求められる．

5　考察

本章では，4 章から得られた分析結果を考察する．4 章で明らかとなった様々な観点からの課題を表 8 にまとめる．1 章では，Peroni らによる定義 [4] に沿って，引用データが「オープン」であることの条件として，構造的（Structured），分離可能（Separable），オープン（Open）であることを挙げた．さらに，引用データの文献が識別可能（Identifiable），入手可能（Available）でなければならないということについても述べた．表 8 では，各観点における課題とともに，オープン・サイテーションの条件のうちどの点において課題があるかを明示するために，構造的，オープン，識別可能という列を設けている．分離可能については，引用データが構造的であれば多くの場合，分離可能でもあるため，省略する．また，入手可能についても，DOI 等によって文献が識別可能であれば，Crossref 等の API を使用すればメタデータを取得することが可能であるため，省略する．表 8 の各課題には記号 A〜G を割り当て，以下でそれぞれについて言及する．

課題 A・G　文献がオープンアクセスであるにも関わらず，引用データがクローズドとなっているケースが観察された．3.1.1 項で述べたとおり，Crossref では，各出版者は各文献の引用データの公開状況を「公開 (open)」「限定 (limited)」「非公開 (closed)」のうちいずれかに設定することができる．Crossref では，2018 年以降に参加した出版者については，デフォルトの状態では引用データが「公開 (open)」になる [30]．それ以前に参加した出版者については，デフォルトの状態では「限定（limited）」となっている [30] が，本稿で利用した引用データのデータセット COCI には「限定（limited）」に設定されている引用データも含まれる．よって，ここで「非公開 (closed)」となっている引用データは，出版者によって意図的に「非公開 (closed)」と設定されたものということになる．近年のオープンアクセスならびにオープン・サイエンスの活動もあって，それらの出版者の中には引用データ公開に関する考えが変化し

表 8: 分析で明らかになった各観点からの課題とその解決策.

観点	課題					解決策
		内容	構造的	オープン	識別可能	
オープンアクセス	A	オープンアクセス誌に掲載されている文献でも，引用データが非公開となっているケースが観察される.	○	×	○	● オープン・サイテーションに関する認知の向上 ● Crossref における引用データの公開設定の変更方法の周知 ● 著者・学術雑誌編集者らによるクラウドソーシングを通じた引用データの登録・オープン化
分類	B	人文学系分野では，引用データが未整理・非存在である文献の割合が高い.	×	○	○	● 機械学習，クラウドソーシング等，引用データの組織化の手法の開発
	C	人文学系分野では引用する文献種別が多岐にわたるが，単行書の章といった学術資料は識別可能ではない.	×	×	×	● 単行書の章等，あらゆる学術資料の識別方法について検討
出版プラットフォーム	D	J-STAGE と大学由来の出版者では，引用データが未整理・非存在である文献の割合が高い.	×	○	○	● 機械学習，クラウドソーシング等，引用データの組織化の手法の開発 ● 引用データの組織化についてノウハウの蓄積・共有
	E	商業出版社では，引用データがクローズドになっている文献の割合が高い.	○	×	○	● オープンアクセスだけではなく，文献に付随するデータのオープン化についても働きかけを実施
出版年	F	過去において，引用データが未整理・非存在である文献の割合が特に高い.	×	○	○	● 機械学習，クラウドソーシング等，引用データの組織化の手法の開発
紀要	G	機関リポジトリで公開されている紀要文献はオープンアクセスであるが，それでも引用データが非公開となっているケースが観察される.	○	×	○	● オープン・サイテーションに関する認知の向上 ● Crossref における引用データの公開設定の変更方法の周知

○はその課題において当該項目には問題がないことを，×は当該項目に特に課題があることを示す.

た出版者も存在するだろう．そのため，Crossref での引用データの公開設定の変更方法といった具体的なアクションとともに，オープン・サイテーションに関する認知を向上させることが必要となる．

また，出版者は引用データを非公開に設定しているが，それに反して著者や学術雑誌編集者が引用データの公開を希望するケースも存在する．このようなケースのために，I4OC の関連組織である OpenCitations は，著者や学術雑誌編集者による引用データのオープン化の方策としてクラウドソーシングを提案しており，クラウドソーシングによる引用データのインデックスとして CROCI（Crowdsourced Open Citations Index）を提供している．CROCI では，著者や学術雑誌編集者等によって作成された CSV 形式や Scholix 形式 [50] の引用データが GitHub を通して OpenCitations に提供されることで，引用データが登録される．CSV 形式では，各行が各引用を示しており，引用文献の識別子，引用文献の出版年月日，被引用文献の識別子，被引用文献の出版年月日から構成される．

課題 B　分類の観点からは，人文学系分野の文献で，引用データのオープン化が STM 分野等の他分野と比較すると遅れていることが判明した．理由として，引用データが未整理・非存在であることが挙げられる．引用データの組織化の方法としては，機械学習や CROCI で採用されているようなクラウドソーシング等の手法が考えられる．

課題 C　人文学系分野の文献では引用する学術資料の種別が多岐にわたることから，あらゆる学術資料を識別可能にする必要がある．代表的なものとして単行本や単行本の章が挙げられる．単行本については ISBN を識別子として利用できる可能性があるが，単行本の章については方策が必要となる．また，文化資料については国立国会図書館デジタルコレクション [51] や新日本古典籍総合データベース [52] における DOI の付与が挙げられ，同様の取り組みが増加することが求められる．

課題 D　出版プラットフォームの観点からの分析では，日本の学術出版物の出版プラットフォームとして最も利用されている J-STAGE と大学由来の出版者の出版プラットフォームでは，引用データが未整理・非存在であることから，引用データのオープン化が進んでいないことが判明した．これらの文献の

多くはオープンアクセスであるので，課題 B 同様，機械学習やクラウドソーシング等の手法を適用することで，引用データの組織化の推進に繋がる可能性がある．また，これらの手法を適用した取り組みを共有することで，リソースが乏しい中でも引用データの組織化を推進するノウハウを蓄積することも求められる．

なお，一般的にリソースが潤沢ではないと考えられる紀要についてであるが，未整理・非存在とラベル付けされた文献は，国立国会図書館分類を使用した分析では 59.74%，文献の出版者名を利用した分析では 47.61%であった（4.7 節参照）．これらの値は，表 2 で提示されている日本の学術出版物における未整理・非存在の割合（73.91%）よりも低い．このことから，引用データをオープンにしている紀要がどのように組織化しているかを共有することは，オープン・サイテーションの推進に貢献すると考えられる．

課題 E　商業出版者の出版プラットフォームでは，引用データがクローズドになっている文献の割合が高いことが観察された．これらの文献ではオープンアクセス化についても課題となっているが，オープンアクセスだけではなく，それに付随するデータ（引用データを含む）のオープン化も働きかける必要がある．1 章で述べたとおり，オープンアクセス出版推進のためのイニシアティブ Plan S では，「I4OC が規定する標準に沿った引用データへのアクセスの提供」を学術雑誌に対して強く推奨している [16]．このような事実を周知することも，オープン・サイテーションの発展に貢献するだろう．また，Crossref は各出版者について参加レポート（Participation Report）(注 4) を作成しており，レポートではその出版者の文献のうち各メタデータ項目がどの程度登録されているかについて示している．この中で引用データについても，登録されている文献の割合とオープンになっている文献の割合が提示されている．

課題 F　出版年という観点からの分析では，日本の学術出版物は，世界の学術出版物と比較すると，過去の文献において引用データのオープン化が進んでいないことがわかった．これは，引用データが未整理・非存在であることが理由である．これらの文献の多くはオープンアクセスであるので，この課題の解決策としては，課題 B 同様，機械学習や CROCI で採用されているようなクラウドソーシング等の手法によって引用データを組織化することが考えられる．

オープンデータは制度面ならびに技術面における再利用性が確保された情報公開の枠組み [53] とされる．オープン・サイテーションはオープンデータの一つの取り組みとして捉えられる．制度面については Crossref で非公開に設定されている引用データを公開することが求められる．オープン・サイテーションについての認知の向上とともに，先述の Plan S のようにオープンアクセスの文脈においてもオープン・サイテーションについて適宜要求していく必要がある．また，各ステークホルダーに対するオープン・サイテーションのインセンティブを明らかにするためにも，学術情報流通に与える影響の分析といった調査も重要となる．技術面については，あらゆる学術資料を識別可能にし，引用データを組織化することが求められる．こちらについては，引用データの組織化についての機械学習やクラウドソーシングといった手法の研究開発を推進するとともに，成功事例をノウハウとして蓄積して共有することが望まれる．

6　おわりに

本稿では，I4OC の活動によって公開された引用データのオープンデータである COCI を利用することで，日本の学術出版物におけるオープン・サイテーションの現状を分析した．世界の学術出版物において引用データをオープンにしている文献の割合は 24.22%であるのに対して日本の学術出版物においては 18.86%であり，海外と比較するとオープン・サイテーションは進んでいない．しかし過去 10 年間に注目すると，日本の学術出版物における引用データをオープンにしている文献の割合は海外と比較して高い．また，日本の学術出版物においては，STM 分野，特にそれらの分野の英文誌において，引用データもオープン化が進展している．対して，歴史・地理，哲学・宗教，芸術等の人文学系分野の文献でのオープン・サイテーションは進んでいない．文献のオープンアクセスの状況は，文献の引用データのオープン化の状況にほとんど影響がないことがわ

(注 4) 参加レポートの例として，https://www.crossref.org/members/prep/286 が挙げられる．

かった．しかし，DOAJ に収録されているオープンアクセスジャーナルにおいては，引用データのオープン化が進んでいる．それでも，オープンアクセスであるにも関わらず，引用データが非公開に設定されている文献が一定数存在しているため，学術情報流通コミュニティにおけるオープン・サイテーションの認知の向上が必要である．

オープン・サイテーションを推進していくには，認知の向上に加えて，引用データの組織化を支援する取り組みが期待される．日本の学術出版物においては，特に過去の文献の引用データが組織化されていないという課題が判明した．過去の文献のオープン・サイテーションを推進するために，機械学習を利用して引用文献リストを抽出し，Crossref メタデータに登録してオープンにするという取り組み [54] がなされている．このように機械と協調することで，過去の文献も含めた学術出版物から効率よく引用データを組織化していくことも求められる．

なお，本稿の分析で作成・使用したデータは，公開されている [55]．

謝辞

本研究の一部は，京都大学東南アジア地域研究研究所・東南アジア研究の国際共同研究「東南アジア地域研究資料のオープン・サイエンス化に向けたとりくみ」，JSPS 科研費若手研究 18K13235 の助成を受けたものです．

参考文献

[1] Smith, Linda C: “Citation analysis”, Library Trends, Vol. 30, No. 1, pp. 83–106, 1981.

[2] Sugiyama, Kazunari; Kan, Min-Yen: “Scholarly paper recommendation via user’s recent research interests”, In Proceedings of the 10th Annual Joint Conference on Digital Libraries, pp. 29–38, ACM, 2010.

[3] “Initiative for Open Citations”, https://i4oc.org/, 最終閲覧日：2019 年 11 月 29 日.

[4] Peroni, Silvio; Shotton, David: “Open Citation: Definition”, 2018.

[5] Tim Berners-Lee: “Linked Data – Design Issues”, https://www.w3.org/DesignIssues/LinkedData.html, 最終閲覧日：2019 年 11 月 29 日.

[6] “OpenCitations”, http://opencitations.net/, 最終閲覧日：2019 年 11 月 29 日.

[7] Wang, Xianwen; Liu, Chen; Mao, Wenli; Fang, Zhichao: “The open access advantage considering citation, article usage and social media attention”, Scientometrics, Vol. 103, No. 2, pp. 555–564, 2015.

[8] Eysenbach, Gunther: “The open access advantage”, Journal of Medical Internet Research, Vol. 8, No. 2, 2006.

[9] Piwowar, Heather; Priem, Jason; Larivière, Vincent; Alperin, Juan Pablo; Matthias, Lisa; Norlander, Bree; Farley, Ashley; West, Jevin; Haustein, Stefanie: “The state of OA: a large-scale analysis of the prevalence and impact of Open Access articles”, PeerJ, Vol. 6, p. e4375, 2018.

[10] “Dimensions”, https://app.dimensions.ai/discover/publication, 最終閲覧日：2019 年 11 月 29 日.

[11] “OpenAIRE”, https://www.openaire.eu/, 最終閲覧日：2019 年 11 月 29 日.

[12] Lauscher, Anne; Eckert, Kai; Galke, Lukas; Scherp, Ansgar; Rizvi, Syed Tahseen Raza; Ahmed, Sheraz; Dengel, Andreas; Zumstein, Philipp; Klein, Annette: “Linked Open Citation Database: enabling libraries to contribute to an open and interconnected citation graph”, In ACM/IEEE on Joint Conference on Digital Libraries, pp. 109–118, ACM, 2018.

[13] Peroni, Silvio; Shotton, David; Vitali, Fabio: “One year of the OpenCitations Corpus”, In International Semantic Web Conference, pp. 184–192, Springer, 2017.

[14] Kacem, Ameni; Mayr, Philipp: “Analysis of search stratagem utilisation”, Scientometrics, Vol. 116, No. 2, pp. 1383–1400, 2018.

[15] Kacem, Ameni; Mayr, Philipp: “Analysis of footnote chasing and citation searching in an academic search engine”, arXiv preprint arXiv:170702494, 2017.

[16] Science Europe: “Plan S – Principles and Implementation”, https://www.coalition-s.org/addendum-to-the-coalition-s-guidance-on-the-implementation-of-plan-s/principles-and-implementation/, 最終閲覧日：2019 年 11 月 29 日.

[17] Xia, Feng; Wang, Wei; Bekele, Teshome Megersa; Liu, Huan: “Big scholarly data: A survey”, IEEE Transactions on Big Data, Vol. 3, No. 1, pp. 18–35, 2017.

[18] Sinha, Arnab; Shen, Zhihong; Song, Yang; Ma, Hao; Eide, Darrin; Hsu, Bo-june Paul; Wang, Kuansan: “An overview of Microsoft Academic Service (MAS) and applications”, In International Conference on World Wide Web, pp. 243–246, ACM, 2015.

[19] Springer Nature: “SN SciGraph”, https://www.springernature.com/gp/researchers/scigraph, 最終閲覧日：2019 年 11 月 29 日.

[20] “DBLP dataset”, https://dblp.uni-trier.de/xml/, 最終閲覧日：2019 年 11 月 29 日.

[21] American Physical Society: “APS Data Sets for Research”, https://journals.aps.org/datasets, 最終閲覧日：2019 年 11 月 29 日.

[22] Tang, Jie; Zhang, Jing; Yao, Limin; Li, Juanzi; Zhang, Li; Su, Zhong: “ArnetMiner: extraction and mining of academic social networks”, In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 990–998, ACM, 2008.

[23] “AMiner Dataset”, https://aminer.org/data, 最終閲覧日：2019 年 11 月 29 日.

[24] Herrmannova, Drahomira; Knoth, Petr: “An analysis of the Microsoft Academic Graph”, D-Lib Magazine, Vol. 22, No. 9/10, 2016.

[25] Harzing, Anne-Wil; Alakangas, Satu: “Microsoft Academic: is the phoenix getting wings?”, Scientometrics, Vol. 110, No. 1, pp. 371–383, 2017.

[26] Haunschild, Robin; Hug, Sven E; Brändle, Martin P; Bornmann, Lutz: “The number of linked references of publications in Microsoft Academic in comparison with the Web of Science”, Scientometrics, Vol. 114, No. 1, pp. 367–370, 2018.

[27] Heibi, Ivan; Peroni, Silvio; Shotton, David: “COCI, the OpenCitations Index of Crossref open DOI-to-DOI citations”, arXiv preprint arXiv:190406052, 2019.

[28] Heibi, Ivan; Peroni, Silvio; Shotton, David: “Crowdsourcing open citations with CROCI-An analysis of the current status of open citations, and a proposal”, arXiv preprint arXiv:190202534, 2019.

[29] 西岡千文; 亀田尭宙; 佐藤翔: “日本の学術出版物における引用データのオープン化の現状分析”, 研究報告人文科学とコンピュータ（CH）, Vol. 2019-CH-120, No. 5, pp. 1–8, 2019.

[30] Crossref: “Members with open references”, https://www.crossref.org/reports/members-with-open-references/, 最終閲覧日：2019 年 11 月 29 日.

[31] Crossref: “Metadata Plus”, https://www.crossref.org/services/metadata-delivery/plus-service/, 最終閲覧日：2019 年 11 月 29 日.

[32] OpenCitations: “COCI CSV dataset of all the citation data (Version 3)”, 2018, https://doi.org/10.6084/m9.figshare.6741422.v3.

[33] OpenCitations: “Download”, http://opencitations.net/download, 最終閲覧日：2019 年 11 月 29 日.

[34] Japan Link Center: “JaLC メタデータ”, https://japanlinkcenter.org/top/material/material_metadata.html, 最終閲覧日：2019 年 11 月 29 日.

[35] “Files for Crossref DOI Dump 201809”, https://archive.org/download/crossref_doi_dump_201809, 最終閲覧日：2019 年 11 月 29 日.

[36] “GitHub – greenelab/crossref: Download metadata for all DOIs using the Crossref API”, https://github.com/greenelab/crossref, 最終閲覧日：2019 年 11 月 29 日.

[37] Unpaywall: “Database Snapshot”, https://unpaywall.org/products/snapshot, 最終閲覧日：2019 年 11 月 29 日.

[38] Darari, Fariz; Prasojo, Radityo Eko; Nutt, Werner: “Expressing no-value information in RDF”, In Proceedings of the International Semantic Web Conference (ISWC) Posters and Demonstrations Track, CEUR Workshop Proceedings, 2015.

[39] 国立国会図書館: “雑誌記事索引”, https://www.ndl.go.jp/jp/data/sakuin/sairokushi.tsv, 最終閲覧日：2019 年 11 月 29 日.

[40] 国立国会図書館: “書誌データ作成ツール：国立国会図書館分類表（National Diet Library Classification：NDLC）”, https://www.ndl.go.jp/jp/data/catstandards/classification_subject/ndlc.html, 最終閲覧日：2019 年 11 月 29 日.

[41] 国立国会図書館: “逐次刊行物”, http://warp.da.ndl.go.jp/info:ndljp/pid/9484238/www.ndl.go.jp/jp/library/data/z.pdf, 最終閲覧日：2019 年 11 月 29 日.

[42] “Directory of Open Access Journals (DOAJ)”, https://doaj.org/, 最終閲覧日：2019 年 11 月 29 日.

[43] Unpaywall: “Data Format”, https://unpaywall.org/data-format, 最終閲覧日：2019 年 11 月 29 日.

[44] Python Software Foundation: “langdetect”, https://pypi.org/project/langdetect/, 最終閲覧日：2019 年 11 月 29 日.

[45] Déjean, Hervé; Meunier, Jean-Luc: “A system for converting PDF documents into structured XML format”, In International Workshop on Document Analysis Systems, pp. 129–140, Springer, 2006.

[46] Tkaczyk, Dominika; Szostek, Paweł; Fedoryszak, Mateusz; Dendek, Piotr Jan; Bolikowski, Łukasz: “Cermine: automatic extraction of structured metadata from scientific literature”, International Journal on Document Analysis and Recognition (IJDAR), Vol. 18, No. 4, pp. 317–335, 2015.

[47] Internet Corporation for Assigned Names and Numbers: “Frequently Asked Questions”, https://newgtlds.icann.org/en/applicants/global-support/faqs/faqs-en, 最終閲覧日：2019 年 11 月 29 日.

[48] Python Software Foundation: “tldextract”, https://pypi.org/project/tldextract/, 最終閲覧日：2019 年 11 月 29 日.

[49] 竹内比呂也: “大学紀要というメディア: 限りなく透明に近いグレイ？（<特集>灰色文献）”, 情報の科学と技術, Vol. 62, No. 2, pp. 72–77, 2012.

[50] Burton, Adrian; Fenner, Martin; Haak, Wouter; Manghi, Paolo: “Scholix Metadata Schema for Exchange of Scholarly Communication Links”, https://doi.org/10.5281/zenodo.1120265, 2017.

[51] 奥田倫子: “国立国会図書館における識別子に関する動向調査”, カレントアウェアネス E, Vol. 2017, No. 332, 2017, https://current.ndl.go.jp/e1950.

[52] 松田訓典; 岡田一祐; 山本和明: “「新日本古典籍総合データベース」にかかわる取り組みと課

題”, じんもんこん 2017 論文集, pp. 219–224, 2017.

[53] 大向一輝: “学術情報流通とオープンデータ（＜特集＞オープンデータ）”, 情報の科学と技術, Vol. 65, No. 12, pp. 503–508, 2015.

[54] Emma Warren-Jones: “Unlocking 100 Years Of Scientific Papers Through Machine Learning”, https://www.scholarcy.com/unlocking-100-years-of-scientific-papers-how-scholarcy-partnered-with-bmj-to-further-i4oc/, 最終閲覧日：2019年 11 月 29 日.

[55] 西岡千文: “日本の学術出版物におけるオープン・サイテーションの分析で作成・使用したデータセット”, https://doi.org/10.14989/244868, 2019.

（2019年 8月 2日 受付）
（2020年 1月10日 採択）